Applied Medical Informatics
Original Research Vol. 40, No. 3-4 /2018, pp: 45-53

Ontology-Based Search Procedure to Identify Tissue Samples in an Autopsy Archive: A Pilot Study

Bogdan POP[1,2*], Bogdan FETICA[2], Mihaiela Luminiţa BLAGA[3], Dan GHEBAN[4,5], Patriciu ACHIMAŞ-CADARIU[6,7], Cătălin Ioan VLAD[6,7], Andrei ACHIMAŞ-CADARIU[1,8]

[1] Department of Medical Informatics and Biostatistics, Iuliu Haţieganu University of Medicine and Pharmacy, 6 Louis Pasteur, 400349 Cluj-Napoca, Romania.
[2] Department of Pathology, The Oncology Institute "Prof. Dr. Ion Chiricuţă", 34-36 Republicii Street, 400015 Cluj-Napoca, Romania.
[3] Department of Information Technology, The Oncology Institute "Prof. Dr. Ion Chiricuţă", 34-36 Republicii Street, 400015 Cluj-Napoca, Romania.
[4] Department of Pathology, Iuliu Haţieganu University of Medicine and Pharmacy, 3-5 Clinicilor Street, 400006, Cluj-Napoca, Romania.
[5] Department of Pathology, Emergency Clinical Hospital for Children, 68 Moţilor Street, 400370 Cluj-Napoca, Romania.
[6] Department of Surgical and Gynecological Oncology, The Emergency Clinical Hospital for Children, 34-36 Republicii Street, 400015 Cluj-Napoca, Romania.
[7] Department of Surgery, The Oncology Institute "Prof. Dr. Ion Chiricuţă", 34-36 Republicii Street, 400015 Cluj-Napoca, Romania.
[8] Department of Internal Medicine, University Hospital C.F.R. Cluj-Napoca, 16-20 Republicii Street, 401167 Cluj-Napoca, România.
E-mails: pop.bogdan21@gmail.com; feticab@yahoo.com; mlblaga@yahoo.com; dgheban@gmail.com; pachimas@umflcluj.ro; catalinvlad@yahoo.it; aachimas@umfcluj.ro

* Author to whom correspondence should be addressed; Tel.: +40 264 598 362;
Fax:+40 264 598 365.

Received: June 25, 2018 /Accepted: August 29, 2018/ Published online: September 5, 2018

Abstract
This study aimed to obtain a detailed record of all autopsy specimens analyzed in the Pathology Department of the Emergency Clinical Hospital for Children, Cluj-Napoca from 1974 to 2018, by using an ontology-based search procedure (OSP) intended to identify the paraffin-embedded stored specimens in pathology reports. Two thousand nine hundred and fifty-six autopsy reports were analyzed using a list of histology terms and expressions commonly found in the microscopic descriptions of the autopsy reports, in Romanian. One pathologist was asked to evaluate the microscopic descriptive part of the autopsy reports for 300 cases and to classify the identified histology specimens according to the ICD-topography codes. The results were then compared with the OSP results. The validation assay returned a 97.32% sensitivity and a 99.48% specificity of the applied ontology-based search procedure when taking as a reference the assessment performed by a pathologist. The most common specimens identified were in the categories of the lower respiratory system (lung, trachea), liver, biliary tract, pancreas and urinary system. The proposed ontology can link valuable information to a highly reliable pathology-based autopsy registry allowing researchers to gain access to specimens stored in the pathology archives, and to facilitate disease registration, data extraction and reporting. This procedure represents a good starting point for developing suitable solutions to be implemented in registries, data banks and for the development of ontology-based registration tools.

Keywords: Autopsy; Child; International Classification of Diseases (ICD); Registry; Archive

Introduction

In the context of widespread use of electronic health records, there is an increasing demand for secondary use of medical data for clinical research, epidemiologic studies or quality assessment studies. A significant barrier is the use of different data models and terminology that are often not interoperable. A considerable amount of data is in the form of unstructured free text that can provide comprehensive information, but requires complex natural language processing methods [1]. For this purpose, increasing use of the semantic web technologies for managing the knowledge of health information systems has emerged [2-4]. The Semantic Web is the next-generation world wide web, in which information is structured and tagged, enabling better cooperation between computers and humans. Ontologies, which can be seen as a more complex collection of terms, constitute the standard knowledge representation mechanism for the Semantic Web [5].

This study aimed to obtain a detailed record of all autopsy specimens analyzed in the Pathology Department of the Emergency Clinical Hospital for Children, Cluj-Napoca from 1974 to 2018, by using an ontology-based search procedure (OSP) intended to identify the paraffin-embedded stored specimens in the pathology reports.

Material and Method

Selection and Description of Participants

In this study, 2,956 autopsy cases, performed in the Pathology Department of the Emergency Clinical Hospital for Children Cluj-Napoca from 1974 to 2018, were retrospectively analyzed. The electronic database of the department is represented by a proprietary database that allows storage and exports of mainly detailed autopsy pathology reports. The information stored in the database does not allow the automatic identification of paraffin-embedded stored specimens in pathology reports, which need to be retrieved for each case. The available data exported from the pathology database were represented by demographical data (such as age, gender, and environment), clinical data (such as hospital admission date, the date of death, the place where the death occurred, clinical causes of death and related diagnoses and co-morbidities) and pathology data (such as autopsy date, pathology causes of death, related diagnoses and comorbidities, description of macroscopic and microscopic lesions).

Building the Ontology

A list of histology terms and expressions, in Romanian language, that are commonly used in the microscopic descriptions of the autopsy reports, was selected in order to identify paraffin embedded stored specimens and to classify them according to the anatomical sites classified by the ICD-0-3 topography (ICD-Topo) codes for each category (C00 to C79) [6]. The list of terms reunites words and expressions commonly used in the microscopic descriptive part of the autopsy reports. The ontology is based on the ICD-O-3 coding systems and has 73 semantic tags (class constraints) corresponding to the topographical codes and 1,844 vocabulary entities consisting of plain keywords or key phrases, varying from 7 to 107 entities per class. A semantic search engine has been implemented in VBA using the classic keyword search paradigm. For each ICD-Topography category common histological terms that are specific for the topographical area were used along with descriptive terms and diagnostic formulations that refer to pathologies that are site-specific (Figure 1).

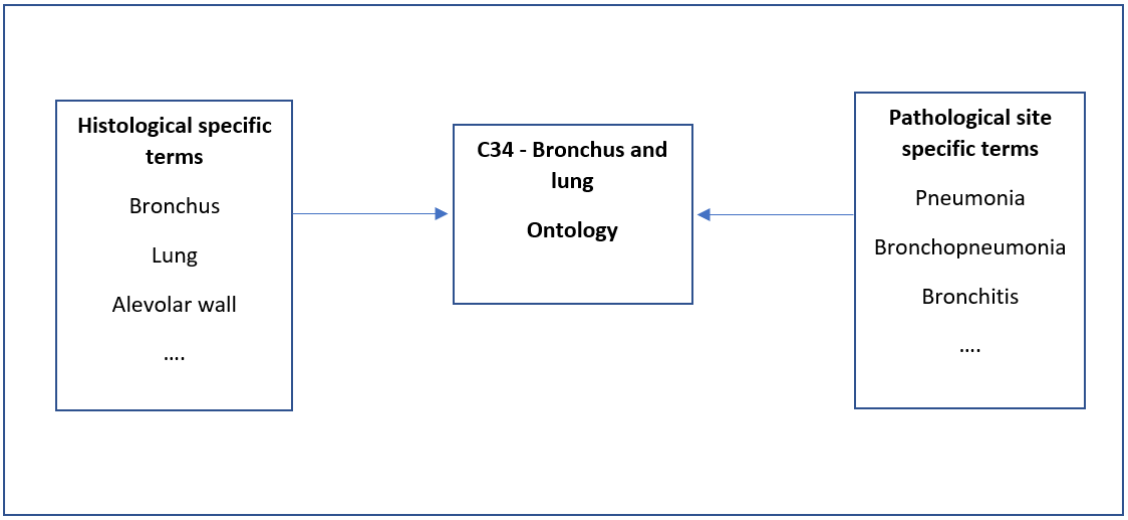

Figure 1. Schematic representation of the construction of the ontology for the C34- Bronchus and lung category

The list was generated by taking into consideration terms that can help identify specific anatomic sites and do not overlap for several anatomical sites. The first version of the ontology was used as source terms for the search module that extracted data from the electronic exports of the autopsy database. Several trial runs allowed the exclusion of overlapping ontology terms (ontology terms that were common to more than one anatomical site) and the addition of several site-specific terms missed in the initial version of the ontology.

Comparative Study of the OSP and the Pathologist's Evaluation

In order to evaluate the final version of the ontology 300 autopsy cases were selected from the 2,832 cases with available microscopic description extracted from the proprietary database, representing approximately 10.5% of the total number of cases. To compensate for changes in terminology and spelling over time (due to the investigated time range), approximately 10 cases for every 100 cases arranged in chronological order were randomly selected. For each of the 300 cases, a pathologist selected, from the microscopic descriptive part of the autopsy reports, the paraffin-embedded stored specimens described for each case and classified them according to the ICD-Topo [6]. The results were then compared with the OSP results.

Statistics

The evaluation performed by the pathologist was considered as reference. In this context, true positive site-matches were the ones identified as positive by both the OSP and the pathologist. True negative site-matches were considered those identified as negative by both the OSP and the pathologist. False positive site-matches were considered the sites identified as positive by the OSP, but classified as negative by the pathologist. False negative site-matches were considered the sites that were identified as negative by the OSP but classified as positive by the pathologist. The assessment of the OSP was performed by calculating the sensitivity (Se) and specificity (Sp) of the assay about the pathologist's assessment by using the formulas:

$$Se = \frac{True\ positive\ site - matches}{True\ positive\ site - matches\ + False\ negative\ site - matches} * 100$$

$$Sp = \frac{(True\ negative\ site - matches)}{True\ negative\ site - matches + False\ positive\ site - matches} * 100$$

Descriptive statistics then summarized the two measures: mean, median, standard deviation, range and 95% confidence interval [7].

Age categories were defined as follows: newborn (0 to 28 days); infant (29 days to <1 year); toddler (≥1year to 3≤years); preschooler (>3years to <6years); schooler (≥6years to 12 years); and adolescent (>12 to 18 years).

Results

Study Group

The extracted data contained only partial demographic information. Data regarding the gender of patients were available for all cases and information regarding the age of the patients were available in the electronic form only for 2,830 cases (95.74%). The study group had a relatively balanced disposition gender-wise with a slight dominance of male patients (54.80%), with a female-to-male ratio of 0.82. More than half the cases with available data in the study group were infants, and more than 75% had less than one year of age at the time of death (Figure 2). The male dominance was more pronounced in infants (56.86%).

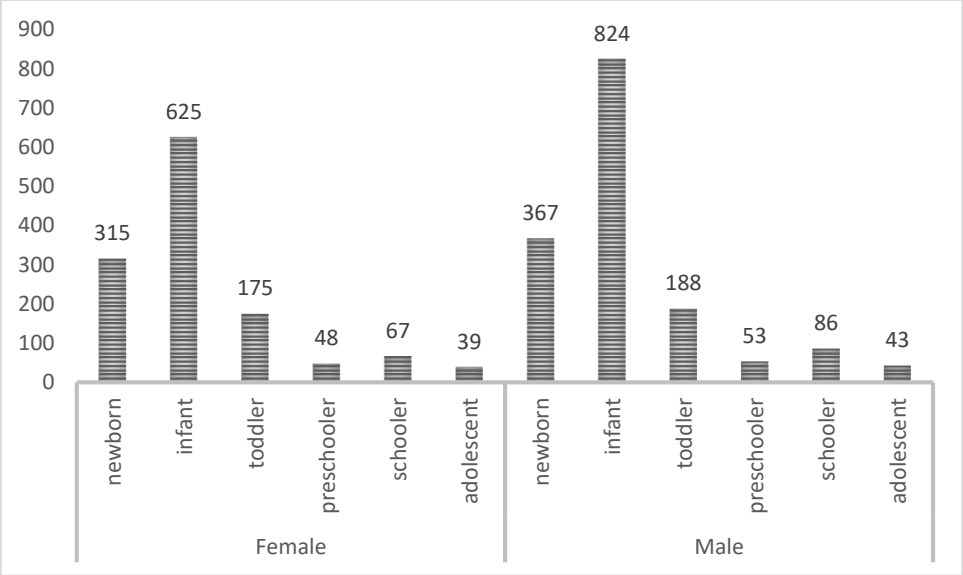

Figure 2. Comparative plots of age groups by gender in the study group

Patient provenance information was available for 2,413 cases (81.38%). Of these more than half (1,360 cases) resided in urban dwellings (56.36%) with the rest of 1,053 cases residing in rural areas (43.64%). Data regarding the county in which patients inhabited were available for 2,006 cases (67.86%). Of these more than 1,761 lived in the NW region of Romania (6 counties: Bihor, Bistriţa-Năsăud, Cluj, Maramureş, Satu-Mare, and Sălaj), representing 87.39%. Another 222 cases lived in neighboring counties from the Transylvania historical region representing 11.01%. Only 1.14% of cases lived in other regions of Romania.

The date of the autopsy was available for all cases. A constant declining trend in autopsy number after grouping the cases in five years intervals was observed (Figure 3). An abrupt decrease in autopsy numbers was observed for the period between 1989-1993.

Comparative Results of the OSP and the Pathologist's Evaluation

For this paper the term "site(s)" refers to specimens identified and classified according to the ICD-Topo. In the validation series of 300 autopsy cases, the pathologist identified a total of 2,208 sites, with an average of 7.36 sites/case and a median of 7 sites/case. The minimum number of sites identified was two sites/case, and the maximum number of sites identified was 21 sites/case. The OSP identified a total of 2,148 sites, with an average of 7.16 sites/case and a median of 7 sites/case.

The minimum number of sites identified by the OSP was two sites/case, and the maximum number of sites identified was 20 sites/case. The software did not identify a total of 60 sites representing 0.02% of the total number identified by the pathologist. Also, the software identified 95 additional sites that were considered false-positive results, which represented 0.043% of the total number of sites identified by the pathologist and 0.04% of the total number of sites identified by the OSP.

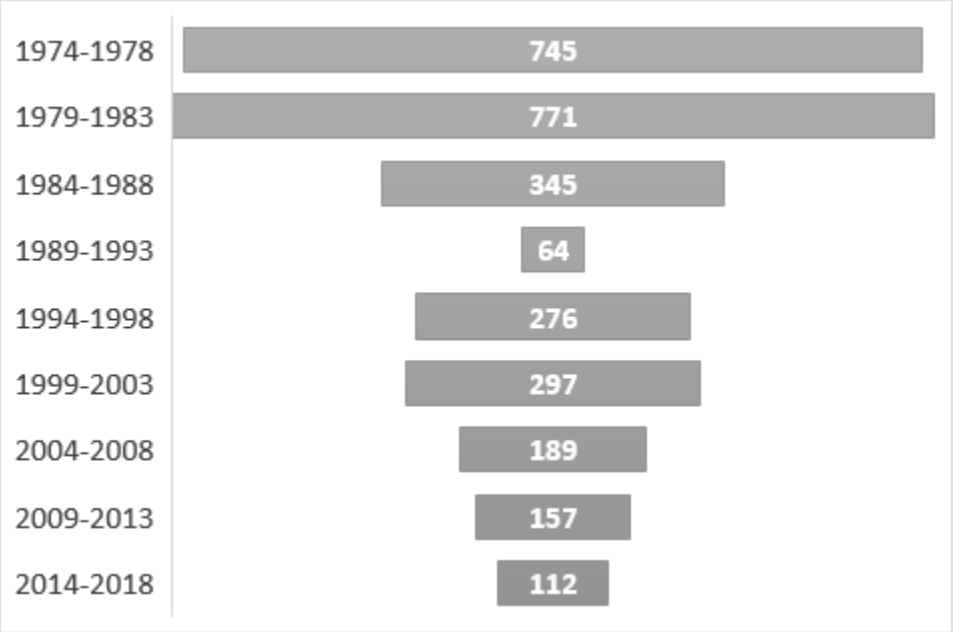

Figure 3. Trends in autopsy rates presented by using five years intervals

By taking into consideration the total number of possible sites/case (codes C00 to C79) and the number of autopsy cases that were assessed, the total possible number of sites was 20700. The calculated Sensitivity was 97.32%, and the Specificity was 99.48% for the OSP compared to the pathologist's results. Due to the substantial number of possible sites, the Sensitivity and the Specificity for each case was calculated, the summarized results being presented in Table 1. When the Sensitivity was assessed for individual cases, a wide range was observed with a minimum of 66% and a high standard deviation (Table 1). Nonetheless, in 249 of 300 cases (83%), the software identified 100% of the anatomic sites. The range and standard deviation for the Specificity were much lower, and in 222 of 300 cases (74%), the Specificity was of 100%. These high values can be attributed to the high number of possible anatomic sites.

Table 1. Summarization for the Sensitivity and Specificity parameters of the OSP assessed for individual cases

Parameter	Sensitivity	Specificity
Count	300	300
Mean	97.32	99.48
Median	100	100
Standard Deviation	6.52	0.99
Range	33.33	5.08
Minimum	66.66	94.91
Maximum	100	100
Confidence interval (95.0%)	96.58 to 98.02	99.37 to 99.59

To have a more informative view of the results a comparison was performed between the number of false-positive sites/case identified by the OSP and the total number of sites identified by the pathologist/case. The mean percentage of false-positive sites was, in this context, of 4.17% with a Standard Deviation of 8.24 and a maximum of 60%/case (i.e., five sites identified correctly by the pathologist and OSP and three supplemental, false-positive sites identified by the OSP). The values should be examined in the light of a highly variable number of true-positive sites/case. Nonetheless, most common false-positive matches were observed for C24 (Other and unspecified parts of biliary tract) and C65 (Renal pelvis) totaling 43 false-positive results (45% of total false-positive). Most of these matches were also accompanied by true-positive C22, C23 and C64 matches (identified by the pathologist and by the software for each case) and were most likely due to similarities in ontology terms for each of these closely related anatomic sites. Unavoidable term similarities represented other examples of false - positive matches (e.g., similarities between C18 ontology term "colon" and the microscopic descriptive term "colonies" which accounted for nine false-positive results (9% of false-positive).

The Ontology-based Search Procedure Assay

Microscopy descriptive data were available for 2,827 cases (93.64%). The OSP was used to identify and classify the examined tissue samples for individual cases. For descriptive purposes, the results were grouped into several categories by using anatomic and pathologic criteria (Table 2). The most common specimens were in the categories of the lower respiratory system (lung, trachea), liver, biliary tract, pancreas and urinary system and at the other end of the spectrum the upper aero-digestive tract (lip, mouth, pharynx), upper airways and the osteoarticular system.

Table 2. Summary of the Ontology-based Search Procedure results

Category	ICD-O-3 Topography	Number of cases	Percent of cases*
Lip, mouth, pharynx	C00-C14	66	2.33
Digestive tract and peritoneum	C15-C21, C26, C48	1448	51.22
Liver and intrahepatic bile duct, gallbladder and biliary tract, pancreas	C22-C25	2653	93.85
Nasal cavity, middle ear, accessory sinuses, larynx	C30-C32	20	0.71
Lung, trachea	C33, C34	2673	94.55
Heart, mediastinum, and pleura	C38	1515	53.59
Thymus, spleen, bone marrow, lymph nodes (hematopoietic and reticuloendothelial systems)	C37, C42, C77	2474	87.51
Bones, joints and articular cartilage	C40-C41	24	0.85
Skin, connective, subcutaneous and other soft tissues	C44, C49	88	3.11
Central, peripheral nerves and autonomic nervous system, eye and adnexa	C47, C70, C71, C72, C69	2019	71.42
Breast and female reproductive system	C50-C58	125	4.42
Male reproductive system	C60-C63	107	3.78
Urinary system	C64-C68	2359	83.45
The thyroid gland, adrenal gland, and other endocrine glands	C73-C75	348	12.31

* Percent of total number of cases with available data

Discussion

The current study presents an ontology-based search procedure (OSP) able to classify the paraffin-embedded specimens sampled and examined from pediatric autopsy cases stored in the archives of the Pathology Department of the Emergency Clinical Hospital for Children Cluj-Napoca between 1974 and 2018. The specimens have been classified according to the ICD-Topo codes for the anatomic location [6]. Two thousand nine hundred and fifty-six pediatric autopsy cases represented the study group, the clear majority from the NW region of Romania. A constant decrease in autopsy rates has been observed in the interval between 1974-2018. The comparison between the assay returned a high sensitivity and specificity of the OSP when taking as a reference the assessment performed by a pathologist. The most common specimens identified were in the categories of the lower respiratory system (lung, trachea), liver, biliary tract, pancreas and urinary system.

When referring to computer sciences, the term ontology refers to a vocabulary that is dedicated and specialized to a particular domain or subject [8]. The term ontology is meant to describe a particular field and the relationship between its parts [8]. Another sense of the term is represented by "the body of knowledge describing some domain, typically a common-sense knowledge domain, using a representation vocabulary" [8]. As the requirement of high-quality data, that can be shared across multiple platforms of healthcare enterprises, becomes a necessity, the next-generation disease registries must enable collection, aggregation and permission sharing of such data [9]. Ontologies can be used to harmonize data across devices and various data sources [9, 10].

The presented search procedure and most of all the list of terms and expression represent a good starting point that allows building upon. A software solution that targets the usage of microscopy descriptive terms in Romanian can take advantages of this list in several applications used for data extraction purposes. An example of such endeavors could be represented by the building of a repository of pathology specimens stored in the laboratories in various regions of Romania. A recognized practical example of such an enterprise is represented by the Danish National Pathology Registry and Danish Pathology Data Bank [11]. The information contained in them have a nationwide coverage of pathology specimens dating back to the 1970s [12]. The Registry and the Data Bank have the benefit of being linked to high-quality population-based registries [11-15].

Although numerous solutions that allow extraction of data from registries are currently available, these are hampered by the lack of common semantics and have limited applicability on data from external sources that require linking [16].

The study reported a constant, significant decrease in autopsy rates in the period between 1974-2018, with an abrupt decrease in the interval between 1989-1993 that in our opinion must be interpreted in the broad context of the socio-political changes, in our country, which has overlapped with significant changes in healthcare and legislation. The current analysis is based on data collected from a single institution, but by taking into consideration the distribution of cases and the fact that the Department of Pathology of the Emergency Clinical Hospital for Children Cluj-Napoca performs autopsies on patients admitted in all the significant pediatric hospital in the city of Cluj-Napoca, we consider the result as being representative at least for Cluj County. Also, when analyzing the distribution of cases, most of the cases are from the North-West region of Romania. The data presented here can be of valuable information in this broader geographical context. It is our opinion that the current trend would be observed in several Pathology Departments throughout the region, but this statement is based solely on empirical and personal observation. Similar trends have been observed in other regions around the world, as requests for pathological autopsies have suffered a dramatic decline over the last 50 years in the United States and countries in Western Europe [17-20].

The results of the search module most likely mirror the abnormal findings in the macroscopic evaluation of individual cases, but the comparison with the case diagnostics must confirm this. One of the future reuses of the results of this study is to link these data to a pathology-based autopsy registry that will ease the access to pathology databases and archived specimens. This recommended modular approach, to building ontology-based registries was chosen in order to allow the use of this module in other related domains [21, 22]. One of the domains for which we intend to extend the use of the OSP is the implementation of a lymphoma registry in our region and the development of a user-oriented ontology-based registration tool suitable to facilitate lymphoma registration in our

region [23, 24]. The high sensitivity and specificity of the OSP can be partially attributed to the fact that it has been "customized" to the currently available data throughout the steps described in the Method part of the paper. The validity of this OSP must be tested on several types of data from various sources and most likely from several regions, in order to have a "real-world" view of the functionality of the system. This initial functional version of the OSP is not exhaustive as it does not include all the possible site-specific histological and pathological terms. Also, the search procedure does not use exceptions (constraints) for specific terms like the ones described above. Further development of this initial version of the OSP will address these issues.

Conclusions

By using an ontology-based search procedure on an extensive series of autopsy cases from the NW region of Romania, we were able to extract and classify, according to the ICD-0-3 topography, the examined pathology specimens for individual cases. Such procedures can be used to build upon suitable for implementation in registries, data banks and the development of ontology-based registration tools.

List of abbreviations

OSP- ontology-based search procedure
ICD-Topo - ICD-0-3 topography

Ethical Issues

The present study was approved by the Ethics Committee of the Iuliu Haţieganu University of Medicine and Pharmacy Cluj-Napoca (Approval number 206/19.04.2018)

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010;17(5):507-13.
2. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC medical genomics. 2011;4(1):13.
3. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics. 2012;13(6):395.
4. Mate S, Köpcke F, Toddenroth D, Martin M, Prokosch H-U, Bürkle T, et al. Ontology-based data integration between clinical and research systems. PloS One. 2015;10(1):e0116656.
5. Zhang H, Guo Y, Li Q, George TJ, Shenkman EA, Bian J. Data Integration through Ontology-Based Data Access to Support Integrative Data Analysis: A Case Study of Cancer Survival. Proceedings IEEE International Conference on Bioinformatics and Biomedicine. 2017;1300-1303.
6. Fritz AG, Percy C, Jack A, Shanmugaratnam K, Sobin L, Parkin D, et al. International classification of diseases for oncology: ICD-O: World Health Organization; 2000.

7. Watson P, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology. 2010;73(9):1167-79.
8. Chandrasekaran B, Josephson JR, Benjamins VR. What are ontologies, and why do we need them? IEEE Intelligent Systems and their applications. 1999;14(1):20-6.
9. Mandl KD, Edge S, Malone C, Marsolo K, Natter MD. Next-generation registries: fusion of data for care, and research. AMIA Summits on Translational Science Proceedings. 2013;2013:164.
10. Spjuth O, Krestyaninova M, Hastings J, Shen H-Y, Heikkinen J, Waldenberger M, et al. Harmonising and linking biomedical and clinical data across disparate data archives to enable integrative cross-biobank research. European Journal of Human Genetics. 2016;24(4):521.
11. Erichsen R, Lash TL, Hamilton-Dutoit SJ, Bjerregaard B, Vyberg M, Pedersen L. Existing data sources for clinical epidemiology: the Danish National Pathology Registry and Data Bank. Clinical epidemiology. 2010;2:51.
12. Nguyen-Nielsen M, Svensson E, Vogel I, Ehrenstein V, Sunde L. Existing data sources for clinical epidemiology: Danish registries for studies of medical genetic diseases. Clinical epidemiology. 2013;5:249.
13. Schmidt M, Pedersen L, Sørensen HT. The Danish Civil Registration System as a tool in epidemiology. European journal of epidemiology. 2014;29(8):541-9.
14. Maret-Ouda J, Tao W, Wahlin K, Lagergren J. Nordic registry-based cohort studies: Possibilities and pitfalls when combining Nordic registry data. Scandinavian journal of public health. 2017;45(17_suppl):14-9.
15. Brooke HL, Talbäck M, Hörnblad J, Johansson LA, Ludvigsson JF, Druid H, et al. The Swedish cause of death register. European Journal of Epidemiology. 2017;32(9):765-73.
16. Esteban-Gil A, Fernández-Breis JT, Boeker M. Analysis and visualization of disease courses in a semantically-enabled cancer registry. Journal of biomedical semantics. 2017;8(1):46.
17. Loughrey M, McCluggage W, Toner P. The declining autopsy rate and clinicians' attitudes. The Ulster medical journal. 2000;69(2):83.
18. Nakhleh RE, Baker PB, Zarbo RJ. Autopsy result utilization: a College of American Pathologists Q-probes study of 256 laboratories. Archives of Pathology and Laboratory Medicine. 1999;123(4):290-5.
19. Thorning D. The role of autopsy in the prevention of medical errors. Laboratory Medicine. 2001;32(5):248-9.
20. Chariot P, Witt K, Pautot V, Porcher R, Thomas G, Zafrani ES, et al. Declining autopsy rate in a French hospital: physicians' attitudes to the autopsy and use of autopsy material in research publications. Archives of pathology & laboratory medicine. 2000;124(5):739-45.
21. Rector A, Brandt S, Drummond N, Horridge M, Pulestin C, Stevens R. Engineering use cases for modular development of ontologies in OWL. Applied Ontology. 2012;7(2):113-32.
22. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology. 2007;25(11):1251.
23. Fetica B, Pop B, Blaga ML, Fulop A, Dima D, Zdrenghea MT, et al. High prevalence of viral hepatitis in a series of splenic marginal zone lymphomas from Romania. Blood Cancer J. 2016;6(11):e498.
24. Fetica B, Achimas-Cadariu P, Pop B, Dima D, Petrov L, Perry AM, et al. Non-Hodgkin lymphoma in Romania: a single-centre experience. Hematol Oncol. 2017;35(2):198-205.